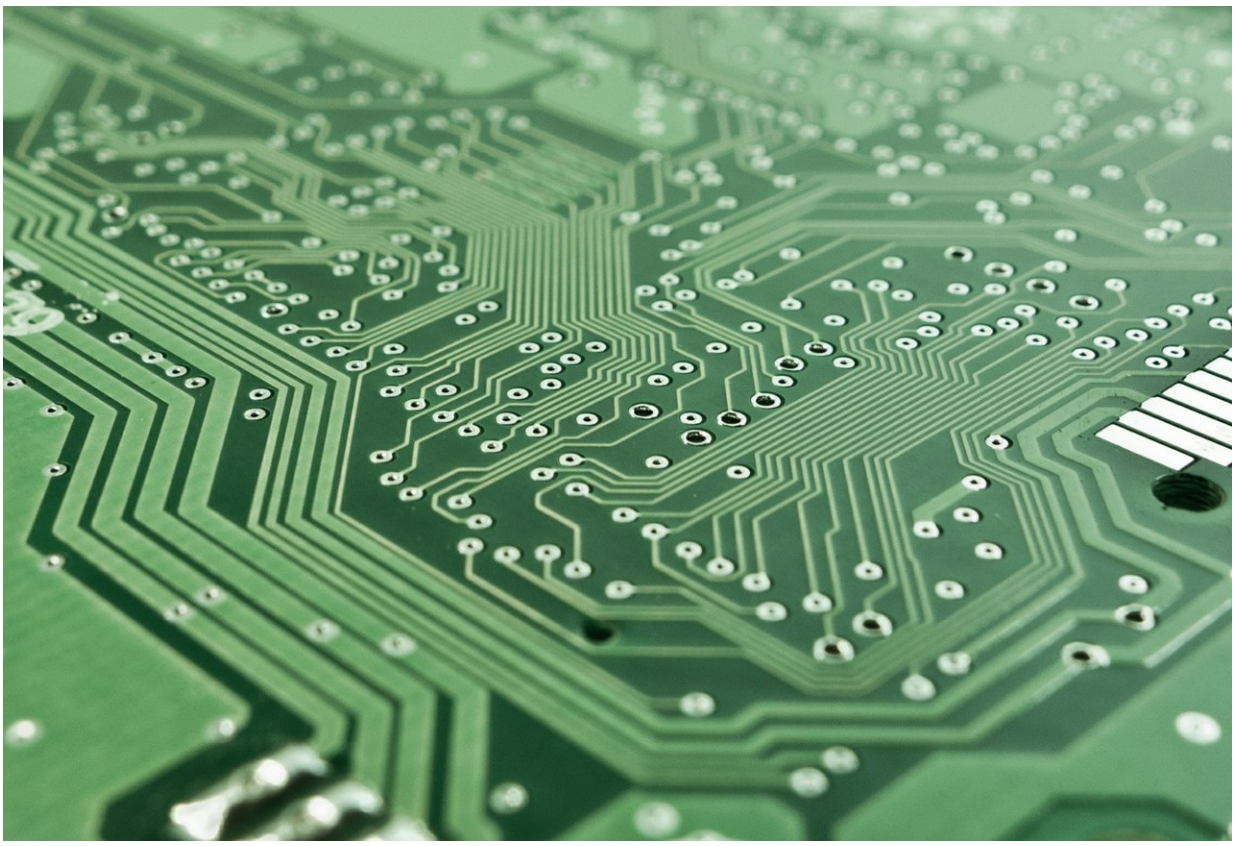
Researchers develop tiny chip that can safeguard user data while enabling efficient computing on a smartphone

April 23 2024, by Adam Zewe



Credit: CC0 Public Domain

Health-monitoring apps can help people manage chronic diseases or stay on track with fitness goals, using nothing more than a smartphone.

However, these apps can be slow and energy-inefficient because the vast machine-learning models that power them must be shuttled between a smartphone and a central memory server.

Engineers often speed things up using hardware that reduces the need to move so much data back and forth. While these machine-learning accelerators can streamline computation, they are susceptible to attackers who can steal secret information.

To reduce this vulnerability, researchers from MIT and the MIT-IBM Watson AI Lab created a machine-learning accelerator that is resistant to the two most common types of attacks. Their chip can keep a user's health records, financial information, or other sensitive data private while still enabling huge AI models to run efficiently on devices.

The team developed several optimizations that enable strong security while only slightly slowing the device. Moreover, the added security does not impact the accuracy of computations. This machine-learning accelerator could be particularly beneficial for demanding AI applications like augmented and virtual reality or autonomous driving.

While implementing the chip would make a device slightly more expensive and less energy-efficient, that is sometimes a worthwhile price to pay for security, says lead author Maitreyi Ashok, an electrical engineering and computer science (EECS) graduate student at MIT.

"It is important to design with security in mind from the ground up. If you are trying to add even a minimal amount of security after a system has been designed, it is prohibitively expensive. We were able to effectively balance a lot of these tradeoffs during the design phase," says Ashok.

Her co-authors include Saurav Maji, an EECS graduate student; Xin

Zhang and John Cohn of the MIT-IBM Watson AI Lab; and senior author Anantha Chandrakasan, MIT's chief innovation and strategy officer, dean of the School of Engineering, and the Vannevar Bush Professor of EECS. The research will be presented at the IEEE Custom Integrated Circuits Conference (CICC), held April 21–24 in Denver.

Side-channel susceptibility

The researchers targeted a type of machine-learning accelerator called digital in-memory compute. A digital IMC chip performs computations inside a device's memory, where pieces of a machine-learning model are stored after being moved over from a central server.

The entire model is too big to store on the device, but by breaking it into pieces and reusing those pieces as much as possible, IMC chips reduce the amount of data that must be moved back and forth.

But IMC chips can be susceptible to hackers. In a side-channel attack, a hacker monitors the chip's power consumption and uses statistical techniques to reverse-engineer data as the chip computes. In a bus-probing attack, the hacker can steal bits of the model and dataset by probing the communication between the accelerator and the off-chip memory.

Digital IMC speeds computation by performing millions of operations at once, but this complexity makes it tough to prevent attacks using traditional security measures, Ashok says.

She and her collaborators took a three-pronged approach to blocking side-channel and bus-probing attacks.

First, they employed a security measure where data in the IMC are split into random pieces. For instance, a bit zero might be split into three bits

that still equal zero after a logical operation. The IMC never computes with all pieces in the same operation, so a side-channel attack could never reconstruct the real information.

But for this technique to work, random bits must be added to split the data. Because digital IMC performs millions of operations at once, generating so many random bits would involve too much computing. For their chip, the researchers found a way to simplify computations, making it easier to effectively split data while eliminating the need for random bits.

Second, they prevented bus-probing attacks using a lightweight cipher that encrypts the model stored in off-chip memory. This lightweight cipher only requires simple computations. In addition, they only decrypted the pieces of the model stored on the chip when necessary.

Third, to improve security, they generated the key that decrypts the cipher directly on the chip, rather than moving it back and forth with the model. They generated this unique key from random variations in the chip that are introduced during manufacturing, using what is known as a physically unclonable function.

"Maybe one wire is going to be a little bit thicker than another. We can use these variations to get zeros and ones out of a circuit. For every chip, we can get a random key that should be consistent because these random properties shouldn't change significantly over time," Ashok explains.

They reused the memory cells on the chip, leveraging the imperfections in these cells to generate the key. This requires less computation than generating a key from scratch.

"As security has become a critical issue in the design of edge devices, there is a need to develop a complete system stack focusing on secure

operation. This work focuses on security for machine-learning workloads and describes a digital processor that uses cross-cutting optimization.

"It incorporates encrypted data access between memory and processor, approaches to preventing side-channel attacks using randomization, and exploiting variability to generate unique codes. Such designs are going to be critical in future mobile devices," says Chandrakasan.

Safety testing

To test their chip, the researchers took on the role of hackers and tried to steal secret information using side-channel and bus-probing attacks.

Even after making millions of attempts, they couldn't reconstruct any real information or extract pieces of the model or dataset. The cipher also remained unbreakable. By contrast, it took only about 5,000 samples to steal information from an unprotected chip.

The addition of security did reduce the energy efficiency of the accelerator, and it also required a larger chip area, which would make it more expensive to fabricate.

The team is planning to explore methods that could reduce the energy consumption and size of their chip in the future, which would make it easier to implement at scale.

"As it becomes too expensive, it becomes harder to convince someone that security is critical. Future work could explore these tradeoffs. Maybe we could make it a little less secure but easier to implement and less expensive," Ashok says.

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology